Semantic Interpretation in the Minimal Commitment Paradigm

1. Motivation

- Why are we doing this
- Why is it a contribution, compared to what's out there
- Although research in NLP began very soon after the development of the first computers, there is a pervasive feeling that progress has not kept up with developments in other areas of computation. [this might not be an appropriate thing to say, since you're not offering a way to address the general problem, but rather another limited approach, and there are in fact any number of limited approaches being used.]

2. Background

2.1 Considerations

- [Need to work from an existing semantic theoretical basis]
- Necessity of domain knowledge for semantic interpretation. Probably no one would disagree that domain knowledge is needed for semantic interpretation (definitional).
- But, more strongly, would claim that any linguistic (syntactic) analysis without domain knowledge is impossible, except in toy (trivial) cases.
- Limited first theoretically, Propositional content only, and then not all of that:
- What is the relationship between syntactic structure and semantic interpretation. Are syntactic predications isomorphic with semantic predications. Would seem not. Depending on lexical insertion, some gerunds encode arguments rather than predicates. *John cut the pie [using a knife]*

The Underspecified Approach / The Minimal Commitment Paradigm

- Whatever we attempt it must be doable on some broad (horizontal) scale, but we can have (vertical) focussing. e.g., the Binding project, Coronary Cath. reports, OncoSem

2.2 Components of semantic interpretation

- Logical vocabulary
- Relational vocabulary
  - attributes vs. relations
- Referential vocabulary
- Attitudinal vocabulary
- Propositional attitude
- ??Propositional content
- Where do these go:
  - tense
  - modality

2.2.1 One way to cut it

- Attitudinal content
- Propositional content

- Relational vocabulary
- Referential vocabulary

2.3 Approaches to Semantic Interpretation

- Alshawi (The Core Language Engine): A comprehensive approach
- Partial / unlabeled analysis: Strzalkowski, Zernik
- Udo Hahn
- in the tradition of Shank and Wilks

3. Overview

3.1 Brings together two strands in NLP research

- The need for more domain knowledge (Bates and Weischedel 1993)
- Shallow parsing [get some early references, along with later ones]

3.2 The domain model already exists

characteristics of the domain model

3.3 Underspecified linguistic analysis

a few general comments

4. System Overview

[[steal from BOSC Report]]

5. Domain Knowledge, UMLS

- Metathesaurus
- Semantic Network

6. Input; allowable text structure

- In general, now
- Special processing in ctx (cortex), for Adam (Conclusions)

7. Determining sentence boundaries and tokenization

- Hearst is a co-author on a paper on determining sentence boundaries; there are other papers on this; can comment on the current method and prospects
- Two levels of tokenization
  - Current primary tokenization; multi-word lexical entries and interaction with the Tagger
  - Secondary (higher leve) tokens; currently addresses locally-defined acronyms; more is planned

8. Lexical look-up and lexical information

- What kind of information is available in the SPECIALIST Lexicon and what is currently used
  - What information could be used in extended SemRep
- Multi-word lexical entries
- Current access to information (generate_variant_info)
- The new system; talk to Lan

9. Category label ambiguity resolution—The Xerox Tagger

- consult_tagged_text (multi-word_lexical entries)

10. Minimal commitment syntactic analysis

- Simple NP, but also other MSU's
- Empirical basis
- Formal characteristics

  - would like to say that it is a cascaded finite state transducer (but you don't know this for sure)
- Description, starting with ANLP 2000
  - mark_boundaries [almost certainly finite state]
  - adjust_boundaries [context sensitive (I think)]
  - segment [formally not needed or finite state]
  - identify_heads [finite state on reversed structure or formally not needed]
  - identify_left_mods [finite state or formally not needed]
- Can note that any "chunking" parser would do; need to find simple noun phrases and a few other types of phrases [list]

discussion

- Boundary word method (Tersmette reference for medical text at least)
- Other references for underspecified syntax and noun phrase extractors
- Relies on the output from a stochastic tagger (Xerox), thus the parser takes as input a list of lexical items assigned category labels.

(1) Algorithm

a. **Assign boundaries to categories**
Boundary categories**:** SomePunc[';', ':', '(', ')', '/']
'not', Auxiliary, Complementizer, Conjunction,Modal, Preposition, Verb

b. **Adjust Boundaries**
Add boundary
- Noun or Adj followed by present participle, not followed by boundary or punctuation
- A non-boundary followed by a determiner or pronoun

- [Aux,Conj, Modal, Verb, ‘:’ get their own MSU
- A non-adjective followed by a comma

Delete boundary (tagger made a mistake)

- Determiner][Preposition

c. **Assign internal labels to noun phrases**
A noun phrase is a phrase containing a noun or adjective as its final member. In these cases label the final element as Head and all other items as Mod.

- Effectiveness; Problems mostly fall into two classes
  - clar resolved incorrectly by the Tagger (not one of these classes maybe)
  - participles
  - sequence of contiguous nouns / adjectives
- Efficiency: very fast, constrained only by Lexical Look-up time

11. Mapping to the knowledge source / MetaMap

12. Semantic interpretation

12.1 Overview

- Interpretation of the Referential Vocabulary
  - MetaMap to the UMLS Metathesaurus
  - Word sense ambiguity
- Interpretation of the Relational Vocabulary
  - Intra-NP analysis
  - Inter-NP analysis
- Argument identification details
  - Domain of arguments
  - Semantic types
  - License for reuse (coordination and relativization)

12.2 Introduction

Is there a contrast in this approach with the traditional view of the relationship between syntax and semantic

- Correspondence rules
- Semantic type compatibility for arguments
- No intercalation of dependencies (no crossing lines)
- No reuse of arguments (without license)
- [general principles of argument identification
- The Minimal Commitment Principle (related to Altruism Avoidance): Don’t do anything that has to be undone. What is done may not be complete, but it is not wrong and does not have to be undone. It’s not the complete truth, but it’s nothing but the truth.
- General principles could (should) be articulated out of the following

- Application of constraints / set intersection
- concentration on argument identification
- Necessity of domain knowledge
- No embedding / no recursion [why]; **String Grammar**

12.3 Details

- Based on underspecified syntactic analysis
- Semantic type compatibility for arguments[slightly higher level categorization for coronary arteries and binding terms; sort this out]
- No intercalation of dependencies (no crossing lines)
- No reuse of arguments (without license) [actually, no reuse of anything without license]
- [general principles of argument identification
  - Direction
  - Domain - Includes direction and indicates whether current MSU is included
  - Prepositional argument cues -

Processing proceeds (from left to right) by first interpreting intra-noun phrase relationships in all simple NP's in the sentence.

Subsequently, inter-NP relations are interpreted as cued by indicators.

12.3.1 Indicators

- What they are (compare to Jackendoff's [Alshawi] correspondence rules)
  - Define a relation between a syntactic phenomenon (word or structure) and a semantic predicate (relation in the UMLS Semantic Network)
- Details of what they look like
  - Argument cues are stipulated
- Note that the application of an indicator rule identifies a potential syntactic predicate that needs to be interpreted semantically.
- Further discussion can include
  - The need for more of them
  - Do they need to be limited according to semantic domain?

In this discussion, need to sort out how these group. e.g., auxiliaries with verbs, gerunds with nominalizations, etc.

Indicators are syntactic phenomena which "indicate" a semantic relationships. Currently these are nominalizations (and other relational nouns), verbs, present and past participles,auxiliaries, and prepositions. (the treatment of present participles/gerunds needs work in implementation.)

Each syntactic indicator is treated similarly, as being a syntactic predicate which has two syntactic arguments. This syntactic predication is then interpreted as a semantic predication, as licenced by the UMLS Semantic Network.

One syntactic indicator is structural (rather than lexical), and this is the modifier-head relation in the simple noun phrase.

12.3.2 Correspondence Rules for Indicators

Stipulate a correspondence between syntactic phenomena and semantic predicates in the UMLS Semantic Network.

12.3.3 Intra-NP Relations

The modifier-head indicator specifies the semantic relationship between a head noun and a modifier immediately to its left.

12.3.4 Left and Right Domain

inter_npu_relations/12 first determines the "left partition" of the sentence, which is the list of structures (reversed) preceding the current indicator. The Left partition for "abcX..." where X is the current indicator is "cba".
The notion of Left and Right Domain is used in order to determine where to look for the left and right arguments. These are figured differently for different indicators.

The left domain of the preposition *of* is the immediately preceding MSU. The left domain of a past participle is the current MSU (since the PastPart may be the last item of this MSU) and all MSUs to the left. The left domain for any other indicator is the list of MSUs to the left (i.e. the left partition).

The right domain for indicator prep is inside the current MSU. The right domain for gerunds (may not be properly implemented) is the current MSU and all following MSUs. While the right domain for any other indicator is anyplace to the right of the current indicator.

12.3.5 Inter-NP Relations

This predicate (inter_npu_relations/12) proceeds through the sentence from left to right, stopping at each potential indicator.

The “raw” algorithm:

- Get a potential indicator
- [If it is a prep, check whether it has already been used (as an arg cue)]
- Get corresponding SemNet information (relation and semtypes)
- Determine left and right domain for this indicator
- Get arguments (apply syntactic constraints)
- Check whether head has already been used (predicate coordination and relativization)
- Check for semantic type compatibility - Potential arguments (as identified by “Get Arguments”) which meet the semantic type requirements for the semantic predicate indicated by this indicator are interpreted as the actual semantic arguments for this predicate.
- Deal with coordination (NP coordination)

12.3.6 Argument Identification (Get arguments)

Both arguments are in the left and right domain for this indicator.

12.3.7 Get arguments of verbs

Surface Object - The SPECIALIST Lexicon provides subcategorization information for verbs. The first simple NP to the right meeting the subcategorization requirements for this verb is taken as a potential surface object for this verb.

Surface Subject - The first NP to the left (not the object of a preposition) is taken as a potential surface subject.

12.3.8 Check for passive

If passive cues are detected, the surface arguments for this verb are reversed.

12.3.9 Get arguments of nominalizations

Either one or both of the arguments of a nominalization my be left unasserted. If both arguments are specified, they can fall into a number of patterns.

(2) a. treatment of aneurysms by/with surgery

b. [PRED] [Prep1 Arg1] [Prep2 Arg2]

(3) a. surgical treatment of aneurysms

b. [Arg1 PRED] [Prep Arg2]

(4) a. surgery for the treatment of aneurysms

b. [Arg1] [Prep-x PRED] [Prep Arg2]

(5) a. surgery is a treatment for aneurysms

b. [Arg1] [aux] [PRED] [Prep Arg2]

(6) a. the (best) treatment for aneurysms is surgery

b. [PRED] [Prep Arg2] [aux] [Arg1]

[What you have to do in describing this is determine how much of this is due to general syntax and how much to subcategorization.

12.3.10 Get arguments of prepositions

Potential right argument is the syntactic object of the preposition. Potential left argument is the first NP to the left in the left domain of this preposition.

12.3.11 Coordination and Relativization

12.3.12 Coordination

Perhaps articulate this specifically in the underspecified domain/shallow parsing

- Coordination can function either at the level of propositional structure and outside propositional structure.
- Outside of propositional structure is discourse structure, at least; there may be other levels
- Clausal (S) coordination is a phenomenon of discourse structure.

- The major/only point of coordination in propostional structure is to license reuse of sentence elements.
- The coordinator licences the reuse.
- NP coordination licences reuse of a predicate (In NP coordination, the two conjuncts must be in "parallel function";
- V & VP coordination licenses reuse of an argument, and the argument must be in "parallel function. cf. relativization.
- What about adverbial coordinators, like *after, prior to,* and *followed by*: The generalization still seems to hold, it's just that the second conjunct must be an NP, but if the first is a verb, then the second is a (predicational) nominalization: *John ran a mile prior to falling* vs *John ran a mile and fell.*
- Kari brought up the question of gapping. Does it fit this paradigm. eg. J*ohn cooked the rice and Mary the fish*. Seems to be a type of predicate reuse, and thus is similar to normal NP coordination, except that there are two conjuncts, rather than one. This is like normal NP coordination in that *John* can be thought of as being coordinate with *Mary* and *the fish* is coordinate with *the rice.* We thus have two coordinate structures, signalled by but one coordinator. However, the general rules are followed: A coordinator occurs in between the conjuncts, and the arguments are in parallel function.

12.3.13 Rules - one way to put it

- In propositional coordination, a coordinator conjoins either predicates or arguments, but not both.
- If arguments are coordinated, they must be in parallel function.
- If predicates are coordinated, they share an argument; it must be in parallel function

12.3.14 Rules - another way to put it

- Metaconditions, informally stated
  - all items must be analyzed
  - No item can be "reused" without a license
- Coordination Rules
  - A coordinator occurs between coordinate items
  - Two items which are coordinate must have exactly the same analysis in propositional structure

12.3.15 License for Multiple Use in V and VP coordination

V and VP coordination involve the reuse of an argument. This involves the notion of "parallel function". That is, the reused argument must have the same function in both predications. A shared (or reused) subject occurs with VP coordination, as in *John kissed Mary and hugged Sue*. Verb coordination involves a shared object (as well as a shared subject), as in *John kissed and hugged Sue*.

12.3.16 Relativization is a type of V or VP coordination

(Normal) Relativization is a type of verb coordination which licenses the reuse of an argument (the Head of the RC); it differs from canonical coordination in that the "conjuncts" need not be in parallel function. In parallel function: *The boy, who likes Sue, hugged Ann.* Not in parallel function: *John kissed the girl who likes Sue.* In either case, the sharing (reuse) is licensed by the relativizer. (There is also structural licensing as in *The boy Sue likes kissed Ann.*)

12.3.17 Implementation

The first clause of the predicate simulates VP coordination with shared subject. The decision is made at the time a subject is being sought for the second verb in the coordinate structure.
Either the conjunction or the relativizer have to appear in between the current indicator and the potential argument to its left. This is accomplished by checking in UpToHead, which is the list of MSU extracted from LeftDomain by predicate up_to_head.
The second clause simulates relativization; it was designed for shared subjects, as in *The RCA is a large vessel which arises from the aorta*. The decision is made at the time a subject is being sought for *arises*. (Shared objects not addressed yet.)

surgery for the treatment of aneurysms provides an example of the Licensing Rules preventing an incorrect multiple interpretation. When *treatment* is reached, both possible arguments (*surgery* and *aneurysms*) will already have been used. The absence of either a coordinator or a relativizer prevents their reuse as arguments of *treatment* and thus this predicate correctly goes uninterpreted.

13. Miscellaneous notes

- Adverbial coordinators seem to require that the right conjunct at least be an NP
- Need to change MinComAn so that adverbs occurring immediately to the right of an NP head are put into a new MSU; currently they are included in the previous NP
- Grice
- West Coast functionalists
- Data Oriented Parsing (DOP)
- Dynamic Conceptual Semantics

% ---------- Find An Adverb Cue ----------

This is an adhoc predicate which is necessary for three reasons.
a. prepositions which can be particles are listed in the Lexicon with label "adverb" (in addition to "preposition").
b. the Tagger often incorrectly treats such a preposition as an adverb
c. if such an item follows an actual adverb, they are both put in the
same MSU, such as " x arises [normally off] the y"

In such a case this predicate is the only way to determine that the following NPU is an allowable argument of "arises"

14. Evaluation

14.1 Test collection

57 MEDLINE abstracts on Parkinsons Disease (parkins)
497 sentences (and fragments)
307 semantic propositions produced
81% correct

extracts from textbooks:
on internal medicine, chapter on eating disorders (zoe1)
206 sentences
104 semantic propositions produced
63% correct

on internal medicine, chapter on eating disorders (zoe3)
114 sentences
82 semantic propositions produced
68% correct

on family medicine, section on eating disorders (zoe2)
96 sentences
48 semantic propositions produced
89% correct

on family practice, section on eating disorders (txbfam)
55 sentences
34 semantic propositions produced
79% correct

Totals

Sentences:

967	total
497	MEDLINE Abstracts
470	medical textbooks

Semantic propositions produced

575	total
307	MEDLINE Abstracts
268	medical textbooks

Percent correct

76%	overall
81%	MEDLINE Abstracts
74%	medical textbooks

Evaluation of the program was conducted on data drawn from the medical research and pedagogical literature. The literature was represented by 57 MEDLINE abstracts on Parkinsons Disease containing 497 sentences (and fragments). A further 470 sentences (and fragments) were drawn from extracts from four medical textsbooks (two each from family practice and internal medicine). The system produced 575 semantic propositions from the 967 total sentences (and fragments) it processed. Each of these propositions was then inspected by hand for correctness and 133 were deemed incorrect (442 correct), for an overall precision score of 76%.

The evaluation was not done on the basis of an independent gold standard. Output of the program was assessed for correctness, but not for completeness. That is, precision was checked, but not recall. [Need to take one sentence and illustrate, the major reasons why propositions are missed. It's probably mostly due to lack of a concept in Meta, lack of a relationship in the Semantic Network, and lack of a correspondence rule.]

The MEDLINE Abstracts were checked by a disinterested physician (Charles A. Sneiderman). The textbook extracts were checked by Rindflesch. Note that results are comparable; textbook results in fact were determined to be lower than MEDLINE results.

14.2 Etiology of errors

14.2.1 General figures

[rough figures]
Out of 575 total propositions produced, 133 were deemed incorrect (442 correct)

67 (50% of 133) - Referential vocabulary / word sense ambiguity
20 (15%) - Indicator (Correspondence) Rules

87 (65%) - Total non-syntax / semantic processing

37 (28%) - Due to syntax / semantic processing

9 (7%) - Multiple / not sure

14.3 Errors not due to syntax

Almost two thirds of the errors are not due to syntax. At least half are due to word sense ambiguity, and an additional 15% are due to errors in the formulation of the indicator rules.

The underlying motivation of this study is to provide some insight into the claim that the underspecified syntactic analysis is rich enough to support useful semantic interpretation. In the following discussion I will point out those incorrect analyses which are crucially (inherently) due to the incomplete nature of the syntactic analysis. I will also give some indication of how prevalent these characteristics are. Although it is obviously not possible to give exact figures, based on the percentages of errors seen in the evaluation, incorrect semantic analyses due inherently to the underspecified syntactic analysis are not terribly frequent. A certain percentage of the errors due to syntax are not inherently due to the underspecified syntactic analysis, but rather are due to implementation errors which can be fixed.

14.4 Referential vocabulary (word sense ambiguity)

[[[At some point in the paper, need a discussion of the fact that we have so far not addressed word sense ambiguity]]]

[As might be expected,] a large number (half) of the errors are due to word sense ambiguity. In a great number of cases, the semantic interpretation would be correct, if the proper sense of the word were available, as in

(7) They may have abrasions on the back of the hand from inducing vomiting.

Hand-LOCATION_OF-Back

The concept “Back” in the Metathesaurus means only the dorsal region of the trunk, which has semantic type ‘Body Location or Region’. “Back of the hand”, if it occurred in the Metathesaurus also would have semantic type ‘Body Location or Region’. Thus if the term were disambiguated the proper interpretation would be available.

In another class of cases, the semantic type for the correct sense of the word in question may not be in the Metathesaurus, and thus no relationship in the Semantic Network currently exists

(8) Abnormalities of neurotransmitter concentrations have been reported in blood and cerebrospinal fluid

Blood-LOCATION_OF-Concentration

Although this interpretation looks promising, in fact, the semantic type of “Concentration” in the Metathesaurus is ‘Mental Process’.
Concentration in the sense meant here....

14.5 Incorrect indicator (correspondence rule)

Fifteen percent of the errors were due to incorrect indicator rules. A large number of these were due to the inclusion of a single incorrect rule which was formulated to apply to the semantic relationship obtaining between a modifier and the following head inside a simple noun phrase. Further the rule was formulated to apply in such cases when the modifer had semantic type ‘Substance’ and the head had semantic type ‘Pathologic Function’.

An initial analysis of constructions like *digoxin overdose* seemed to support a rule in which the modifier head relationship in these cases mapped to the Semantic Network Relationship CAUSE, as in

(9) Substance - CAUSES - Pathologic Function

And that therefore the interpretation of *digoxin overdose* was that the digoxin caused the overdose. However, further examples (and reflection) indicate that this is not the correct analysis. A few such errors, like (10), were due solely to this rule.

(10) The mechanism is thought to be estrogen deficiency

Estrogens-CAUSES-deficiency

[[The correct analysis for these constructions should be ....

However, others were due to a combination of factors, including, often, word sense ambiguity as well as the infelicitous rule, as in (11), *consumption* was wrongly interpreted as tuberculosis.

(11) Registered dietitians often function to educate patients on nutritional needs and food consumption

Food-CAUSES-CONSUMPTION <2>

14.6 Errors due to syntax

The underlying motivation of this study is to provide some insight into the claim that the underspecified syntactic analysis is rich enough to support useful semantic interpretation.
In the following discussion I will point out those incorrect analyses which are crucially (inherently) due to the incomplete nature of the syntactic analysis. I will also give some indication of how prevalent these characteristics are. Although it is obviously not possible to give exact figures, based on the percentages of errors seen in the evaluation, incorrect semantic analyses due inherently to the underspecified syntactic analysis are not terribly frequent.
A certain percentage of the errors due to syntax are not inherently due to the underspecified syntactic analysis, but rather are due to implementation errors which can be fixed.

[[[Try to find some examples which are solely due to syntax, or else change the figures above to represent the true etiology. In fact it looks like very few of the errors are inherently due to the underspecified syntax
Such errors as occur involve the “arguments” of prepositional indicators

(12) A calm but realistic review of the dangers of **starvation**, including **sudden death**, should be given, coupled with statements like "my job is to help you deal **with this illness** so that you can have a normal life expectancy **with reasonable happiness**."

Death, Sudden-CO-OCCURS_WITH-Illness, NOS
Happiness-PROCESS_OF-Starvation

All Metathesaurus concepts for the above sentence:
Calmness (fndg)
Realism (idcn)
Review [Publication Type] (inpr)
Dangerousness (idcn)
Starvation (patf, sosy)
Including (ftcn)
Death, Sudden (dsyn)
Occupations (ocdi)
Illness, NOS (patf)
Normal (qlco)
Percent normal (qnco)
Life Expectancy (grpa)
Happiness (menp)

In (13) *consciousness* is wrongly taken as an argument of *resulted*, rather than *Vomiting*

(13) Vomiting during a state of decreased consciousness has resulted in aspiration pneumonia

Pneumonia, Aspiration-RESULT_OF-Consciousness

Vomiting (sosy)
Decreased (ftcn, qlco, qnco)
Consciousness (menp)
Pneumonia, Aspiration (dsyn)

14.6.1 Other arg problems, independent clause

References

Charniak, Eugene, and Yorick Wilks (eds.) (1976) Computational Semantics: An introduction to artificial intelligence and natural language comprehension. Amsterdam: North-Holland.

Schank, Roger (ed.) 1975. Conceptual information processing. Amsterdam: North-Holland.

Strzalkowski, Tomek, and Barbara Vauthey. 1992. Information retrieval using robust natural language processing. Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, 104-111.

Wilks, Yorick. (1972) Grammar, meaning and the machine analysis of language. London: Routledge & Kegan Paul.

Zernik, Uri. 1992. Corpus-based thematic analysis. In Paul S. Jacobs (ed.) Text-based intelligent systems, 101-121. Hillsdale, NJ: Lawrence Erlbaum Associates.

Alshawi, Hiyan (ed.) 1992. The core language engine. Cambridge, MA: The MIT Press.